DOCUMENT RESUME

ED 105 034 UD 015 024

AUTHOR Crain, Robert L.; York, Robert L.
TITLE Evaluation with an Experimental Design: The Emergency School Assistance Program.
INSTITUTION National Opinion Research Center, Chicago, Ill.
PUB DATE 74
NOTE 75p.; Some pages may reproduce poorly due to quality of original document

EDRS PRICE MF-$0.76 HC-$3.32 PLUS POSTAGE
DESCRIPTORS Academic Achievement; Control Groups; Data Analysis; Educational Research; *Evaluation Methods; *Experimental Groups; Experimental Programs; Federal Programs; Matched Groups; Measurement Techniques; *Program Evaluation; *Research Design; Research Methodology
IDENTIFIERS *Emergency School Assistance Program

ABSTRACT

The Evaluation of the Emergency School Assistance Program (ESAP) for the 1971-72 school year is the first application of full-blown experimental design with randomized experimental and control cases in a federal evaluation of a large scale program. It is also one of the very few evaluations which has shown that federal programs can raise tested academic achievement. Finally, it demonstrated that motivational factors and what are sometimes called noncognitive variables are an important part of the analysis of what happens in schools. A block of ESAP funds was awarded on a random basis to pairs of schools, one member of each pair receiving no funds and serving as control on the other. At the end of the school year, students in both the experimental and control schools received questionnaires and achivement tests; black male high school students were found to score significantly higher in the experimental schools than in the controls. This experience provides virtually the most convincing data that scientific research can provide that the program had a favorable impact on student test scores. An analysis of the data using an elaborate multiple correlation and regression design was also performed. This analysis, combined with the experiment, gave some useful insights into why ESAP was a success. (Author/JM)

ED105034

EVALUATION WITH AN EXPERIMENTAL DESIGN: THE EMERGENCY SCHOOL ASSISTANCE PROGRAM

Robert L. Crain*
Robert L. York**

National Opinion Research Center

The Evaluation was done by the National Opinion Research Center (NORC), a not-for-profit institute loosely affiliated with the University of Chicago. NORC is a survey research organization which mainly carries out survey work; it has, however, also done research in a variety of fields, including studies of health, housing, crime and religion as well as education. The support of the National Institutes of Mental Health, Center for Metropolitan Problems is also acknowledged.

---

* The Project Director, Robert L. Crain, was, at the time this study was done, associate professor of sociology at the Johns Hopkins University. He now is with The Rand Corporation, Santa Monica, California. He also has taught at the University of Chicago, where he completed his graduate studies. He is co-author of three books on school desegregation.

** Robert L. York was project monitor for the Office of Education, where he has been with the Office of Planning, Budgeting, and Evaluation for five years. His graduate training was at Rutgers University in sociology. This article was co-authored by Robert York in his private capacity, therefore no official support or endorsement by the Department of Health, Education and Welfare is intended or should be inferred.

UD 015024

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

INTRODUCTION

The Evaluation of the Emergency School Assistance Program (ESAP) for the 1971-72 school year[1] is, we believe, a success story. It is the first application of full-blown experimental design with randomized experimental and control cases in a federal evaluation of a large scale program. But this was only part of the success. It also is one of the very few evaluations which has shown that federal programs can raise tested academic achievement. Finally, it demonstrated that motivational factors and what are sometimes called non-cognitive variables are an important part of the analysis of what happens in schools.

This is the story of that evaluation. Perhaps, because the authors are so pleased with themselves, we are willing to be rather candid about the blunders and foolishness which seem to be an inevitable part of the evaluation process. The value of the pudding now proved, we can cheerfully admit that we had our thumb in the mixing bowl more than once in the process.

Told briefly, the evaluation and its outcome are a simple story. A block of Emergency School Assistance Program funds was awarded on a random basis, to pairs of schools, one member of each pair receiving no funds and serving as control on the other. At the end of the school year students in both the experimental and control schools received questionnaires and achievement tests, and black male high school students were

found to score significantly higher in the experimental schools than in the controls. This experience provides virtually the most convincing data that scientific research can provide that the program had a favorable impact on student test scores.

But this simple story can be elaborated. The story of the evaluation is simple; but the story of how the evaluation was done is more complex. The research team, in part because of its academic bias that the experiment would show no positive effects and that federal programs such as ESAP were unlikely to succeed, created an elaborate analysis using multiple correlation and regression techniques. One thing they learned is how much more clumsy and inadequate regression is, compared to the precision and elegance of an experiment. But this analysis did have a useful result; for combined with the experiment, it gave us some useful insights into why ESAP was a success. This is important, expecially since ESAP has since been superceded by other programs; policy-makers need to know how these favorable effects can be transferred to other programs. Thus what started out as a mistaken strategy--to anticipate the failure of ESAP and to ignore the power of the experimental method--in the long run helped the evaluation.

PART I: THE CONTEXT OF THE EVALUATION

A. THE PROGRAM

Following years of pressure from the Federal government and litigation, many court orders requiring extensive desegregation to break up de jure segregated school systems in the South took effect starting in the 1970-71 school year. Between 1968 and 1970, school districts in 11 Southern States began to be more desegregated than the Northern and Western States. In this period the per cent of black students attending 80-100 per cent minority schools decreased dramatically from 79 per cent to 30 per cent in the South but remained constant--and more segregated by 1970 at 57 per cent--in the North and West.[2] Some additional desegregation occurred the following school year from the precedent established in the U.S. Supreme Court decision in Swann v. Charlotte-Mecklenburg Board of Education. By the Fall of 1971 the per cent of black students attending all-minority schools was slightly greater in the North and West than in the South.[3]

This focused attention on the issue of federal aid to assist desegregating school districts and in late summer and early fall of 1970 grants were awarded by the U.S. Office of Education to school districts under the new Emergency School Assistance Program (ESAP).

By the summer of 1971 when this evaluation was planned, there was little in the record of the ESAP story to entice social researchers. A "quick and dirty" evaluation of the first year of the ESAP program had been completed.[4] While

comparable in quality to much academic and applied research, the contractor was not given a tightly designed study by the evaluation unit in the Office of Education and the analysis of the survey data collected was not very elegant. One volume of case studies (with two bulky appendices) is useful in that it fulls a gap in the literature on details of specific programs to implement school desegregation. But like most case studies--especially those in three volumes--this one attracted little attention.

While ESAP in principle was to provide the financial resources to permit effective desegregation, in practice it looked like little more than a miniaturized program of federal aid to southern schools. The total amount of money awarded, sixty-four million dollars, was spread over several hundred school districts so that each school received less than $10,000. In order to qualify the school district had to be in the process of desegregation and had to present a proposal indicating how it wanted to use the funds. The proposals varied from eloquent to semi-literate and requested monies for everything from remedial teachers to band instruments.[5] A group of organizations concerned with civil rights had, during the first year of ESAP, prepared an impressive report[6] charging that large numbers of grants went to districts engaging in serious and widespread discrimination and generally reminding us that ESAP programs were being designed by the school people who had brought us taken integration.

If there was anything to be excited about in the ESAP program, it was the Washington staff. Headed by Herman Goldberg, an ex-school superintendent who became nationally famous in Rochester for his efforts at desegregation, the Washington staff was liberal and persistent. Perhaps they could take the limited control they had over a small amount of federal aid and work wonders with it, but it did not seem very likely.

At the time our ESAP evaluation was underway, there were harbingers of a major conflict within the Office of Education. We may roughly characterize the contestants as being supporters of the cognitive versus the humanistic schools of educational improvement. The cognitive faction had over the years been relatively successful in earmarking federal aid to programs directly designed to improve cognitive test scores. The non-cognitive faction included many integrationists and others who say the quality of race relations in southern schools as the critical issue. In short, it had many elements of the classic desegregation versus compensatory education debate.

B. THE DECISION TO RANDOMIZE

The evaluation was conducted not by the ESAP staff, but by a special division within the Office of Education--the elementary and secondary school group of the Office of Planning, Budgeting, and Evaluation. Within this group the task of evaluating ESAP from its start had been assigned to Robert

York. During one of the few reflective moments allowed to a federal evaluation researcher, York reread the often quoted, seldom followed recommendations about experimental evaluations made by Campbell and Stanley[7] and decided that ESAP could be evaluated with a genuine randomized experimental design. Anyone familiar with federal evaluation policy will recognize this as a genuinely radical--and perhaps utopian--decision. A randomized design had never been used in a major Office of Education study. Evidence is on record that OE was well aware of what they might be walking into. Donald Campbell had written one of his many papers[8] arguing that randomization experiments were the only acceptable way to do evaluation; one of the two rebuttals to the paper,[9] arguing that one couldn't put all their eggs in the randomization basket because of the enormous political and programmatic obstacles, was coauthored by

John Evans, head of the Office of Education evaluation group. In the rebuttal, Evans wrote "Our experience leads us to conclude, though reluctantly, that in the actual time-pressured and politically loaded circumstances in which social actual programs inevitable arise, the instances when random assignment is practical are rare; and the nature of political and governmental processes makes it likely that this will continue to be the case." But Evans was deeply committed to establis[illegible]quality evaluation; he had been one of the leaders of the Office o[illegible]mic Opportunity's evaluation group which had been a mo[illegible]e rest of the federal government. Evans presented the rationa[illegible] the design (and the weaknesses of alternative designs) very articulately to Goldberg and his top staff. Goldberg was more familiar with and respectful towards research than most program administrators, having had internal research done for him on the desegregation programs he had instituted in Rochester. After consulting with his staff and recognizing the problems that this design could create, Goldberg made what Evans later called a "courageous decision" to proceed with the randomized design. Courage is the important factor here; for the barriers to randomization in this case were political, not technical. Hopefully the ESAP experience will set a precedent which will make randomization easier in the future.

The idea of allocating ESAP funds randomly upset the ESAP staff. Some could hear the complaints coming to Washington from 100 southern school districts. More importantly, some

were deeply concerned that deserving and needy schools would be deprived by a flip of a coin. Fortunately, in this case Evans and York could argue that since the total amount of funds was constant, we were not so much taking money away from schools as we were transferring it from one school to another. Without the randomized design each school district would spread its small amount of federal funds among a certain number of schools. With control schools randomly designated, the districts would have the same amount of fund, but would concentrate it on a slightly smaller number of schools or extend the program to additional schools. Since the amount of funds was in nearly every case small one could argue that the money would still be put to good use in the other schools in the district. Districts with fewer than four schools intended to receive ESAP funds were excluded from the sample to avoid possible problems in reallocating ESAP funds from the control school to only three other schools in the district.

C. THE EXPERIMENTAL DESIGN

The basic plan that York had developed was that all districts which qualified for funds would receive them. but that a sample of the districts (both renewals and new awards) would be drawn and the superintendent there asked to group the recipient high schools and elementary schools into matched pairs. These pairs would then be randomly sampled and one of the two schools in the selected pairs randomly designated as a control school to receive no funds. Between one and six pairs were selected depending on school district size so as

to insure that fewer than one-fourth of the eligible schools in any school district would be designated as control schools. This would produce a sample of 150 experimental and 150 control schools; in order to enlarge the sample for a conventional cross-sectional analysis, data was collected from an additional 300 schools receiving ESAP Funds.

The evaluation staff was more than slightly nervous as they began trying to obtain the cooperation of school superintendents. It is true that ESAP regulations specified that schools must cooperate in an evaluation; however, it said nothing about anything so radical as an experimental design. Furthermore, the sample was drawn just after the first of several batches of grants were awarded, so some of the early grantee's plans were well advanced by the time they were notified of the control school design (this problem was reduced in the remaining grants by including an advance warning in the telegrams announcing the grant awards). If a superintendent decided to tell the Office of Education to jump off the 14th Street Bridge, the staff exerted a modest amount of pressure, although hopefully not enough to cause him to call his Congressman. When the paper blizzard of telegrams and letters was finished, Eugene Tucker, who was in charge of the operation, had managed to lose only 40% of the school districts. This is a small fraction, given the novelty of the experiment and the political problems that control schools might cause for a superintendent. Since these withdrawals occurred before ramdomnization, it does not produce a bias in

the experiment, although it does limit the study's generalizability to the more cooperative 60% of southern districts.[10]

Meanwhile York has been developing the work statement for the Request for Proposal (RFP) and in mid-September the RFP was issued by the Office of Education. The work statement summarized the major objectives of the study as:

> To conduct a program effectiveness evaluation of the ESAP program itself which will focus on (a) achievement test score effects and other measures of achievement related behavior and attitudes and (b) attitudinal and behavioral effects of minority and majority group students and teachers toward each other. This will involve the administration of questionnaires which include but will not be limited to standardized achievement tests and a series of carefully developed measures of the attitudes and behavior of minority and majority group students and teachers toward each other. The Office of Education is now drawing a random sample of ESAP and control schools for the purposes of this evaluation.
>
> To conduct a study of the larger process of school desegregation apart from ESAP by examining the relationship between a large array of student, teacher, school and school district variables believed to be related to effective desegregation and the outcome variables (achievement test scores and attitude and behavior measures referred to in (a) above. This analysis will be independent of the ESAP analysis in the sense that data from ESAP and non-ESAP control schools will be pooled.

In order to conduct this study, questionnaires will be administered to principals, teachers and students. A data collection guide will be prepared to obtain necessary information from the local ESAP project directors.

D. THE NORC PROPOSAL

Seven proposals were received. Four were rated as unacceptable by the review panel. After negotiations with the remaining three, the National Opinion Research Center (NORC) was selected, not so much because of its proposal, which was hastily written and showed more interest in desegregation than in ESAP, but because of the experience of the Study Director (Crain) in school desegregation and survey research and NORC's generally good reputation for its surveys.

The NORC proposal was the only one to come from an academic research institute, and the only one staffed by a bona-fide academic social scientist. Academic groups do not often bid on evaluation contracts. However, York's request for proposal had explicitly called for social scientists with experience in school desegregation, and Crain, in the course of serving on an earlier proposal review committee for OE had discovered how thinly staffed many of the for-profit research firms were. Crain guessed the odds on a NORC proposal including some experienced academics would be

rather favorable; without a certain amount of experience with OE it is unlikely that he or NORC would have prepared a bid. Students of formal organization often point out how lines of communication outside of formal bureaucratic channels increase organizational efficiency, and this seems to be an example.

One reason why academic institutes are reluctant to bid on evaluations is that they can afford to be "choosy" and don't like to waste time writing high-risk proposals. The other reason is that they don't have staff. Academic researchers usually won't put up with being commandeered into a research project outside their area of expertise. NORC has a very small stable of resident researchers, and could not possibly have done the project with its in-house staff. In this case, it put together a complicated coalition based on two ex-employees, Crain at Johns Hopkins University (Baltimore) and James J. Vanecko at Brown (Providence). The proposal called for a small amount of work by James A. Davis; the remaining staff was Carol Stocking, a survey researcher with no academic credentials but good experience, and four graduate students, Jean Jenkins and Terrence Halliday at Chicago, and Janet Griffith and Ruth Narot of Johns Hopkins. The staff was not interdisciplinary; all were sociologists.

Staffing a project with a non-resident study director, a non-resident assistant director, and no one full-time in Chicago who had ever published a data analysis, was painful; it seemed likely that the prime beneficiary of the project might turn out to be the airlines.

In most cases, when a government agency signs a contract with an academic researcher it receives both more than and less than it bargains for. It may receive good quality work, but it may also get a less-than-responsive report, and it may be submitted late. The problem is that money alone is usually not enough to entice an academic into the contract research game--there are easier ways to make a living, and teaching undergraduates is one of them. In this case Crain wanted two things in addition to money; one was the opportunity to create a data tape for a "second generation Coleman Report"; the other was a chance to campaign for the use of non-cognitive measure of school performance to replace the widespread reliance on cognitive tests.

E. AN OVERVIEW OF THE STUDY AND ITS OUTCOME

Crain, Vanecko and NORC agreed that the experimental design, heroic though it might be, was completely uninteresting. If anything had been firmly established in previous evaluations, it was that small dollops of federal funds would do little to change the quality of education. There was, in their mind, not the least chance that ESAP would have an effect.

The design called for the awarding of ESAP funds early in the 1971-72 school year to experimental schools in each randomly selected pair. There would be no pretest; the randomization eliminated the need for one. As Campbell and Stanley[11] write: "For psychological

reasons, it is difficult to give up "knowing for sure" that the experimental and control groups were "equal" before the differential experimental treatment. Nonetheless, the most adequate all-purpose assurance of lack of initial biases between groups is randomization." In the fall, the treated and untreated schools would be identical within the limits of sampling errors.[12] In the spring, when the school outcomes are measured, the evaluation would determine if the "treated" and "untreated" schools were statistically different from each other. With only 50 treatment-control high school pairs and with only 100 elementary school pairs, the experimental schools would need to experience considerable gain in achievement in order for significant differences to appear. Thus if by some miracle ESAP was not a total waste of money, a program wherein funds arrived sometime after school started in the fall and which required an evaluation in May of the same school year hardly deserved the embarrassment of the unavoidable negative evaluation which would result. Equally important, none of the NORC staff had experience with experimental design; they tended to be interested in what they knew best, namely multivariate analysis of cross-sectional data.

The basic idea of the experiment is simple. If the experimental and control schools are selected from the same population by a random process, it is quite unlikely that the two groups will be different; probability theory tells us exactly how unlikely. Thus, when at the

end of the school year we found that black male high school students in the experimental schools were performing one-half grade higher than those in the control schools, we knew that there was less than one chance in twenty of any group of students differing this much between the two groups of schools by pure chance, and were virtually forced to conclude that the program had a positive effect.

One can understand the enormous advantage of an experiment by realizing all of the statistical tools that have been developed to artificially match treated to untreated groups. Cross tabulation with control variables, multiple correlation, multiple regression, standardization and analysis of covariance are all techniques developed to statistically match two groups. The problem is that none of these techniques work very well. We can prove that these techniques must in principle have some error; in this research we think we have some empirical evidence that in the real world of evaluation research the error can be quite large.

The "conventional wisdom" among social scientists was that not only was it a certainty that ESAP was a waste of money, but there was no chance of <u>any</u> of the many projects practiced with ESAP funds being effective. Crain generally took the view that this

"conventional wisdom" that interventions could not affect the quality of education was sharply overstated. Even though it could be taken for granted that the experiment would fail to show that ESAP on the whole had improved the schools, he wanted to be able to show that some of the ideas being practiced with ESAP funds made sense. In short, a method was needed to rescue the successful minority of ESAP projects from being thrown out just because the overall program was worthless. Beginning with this idea, a combination study, using both multiple regression and the experimental design, was developed.

The combined analysis consisted of four questions, as shown in Figure 1. Each question is represented by a arrow head: the two solid lines refer to questions which can be answered by the experiment, the two dotted lines to the cross-sectional regression analysis. The experimental design could tell us two things. First it could tell us whether the ESAP funds actually led to the establishment of new educational programs and resources in the school. If the experimental schools had significantly more remedial reading than the control schools then we would know that ESAP funds were spent on this. Secondly, the experiment could measure the overall effect of ESAP by determining whether the experimental schools were or were not different from the control schools when the experiment was over. But beyond this the regression analysis would have to rescue the expected negative findings. After we had found that the experimental and control schools were not significantly different, we planned to carry out a careful analysis of the impace of all the various school programs, resources, or activities on both cognitive and non-cognitive outcomes, giving us a chance to

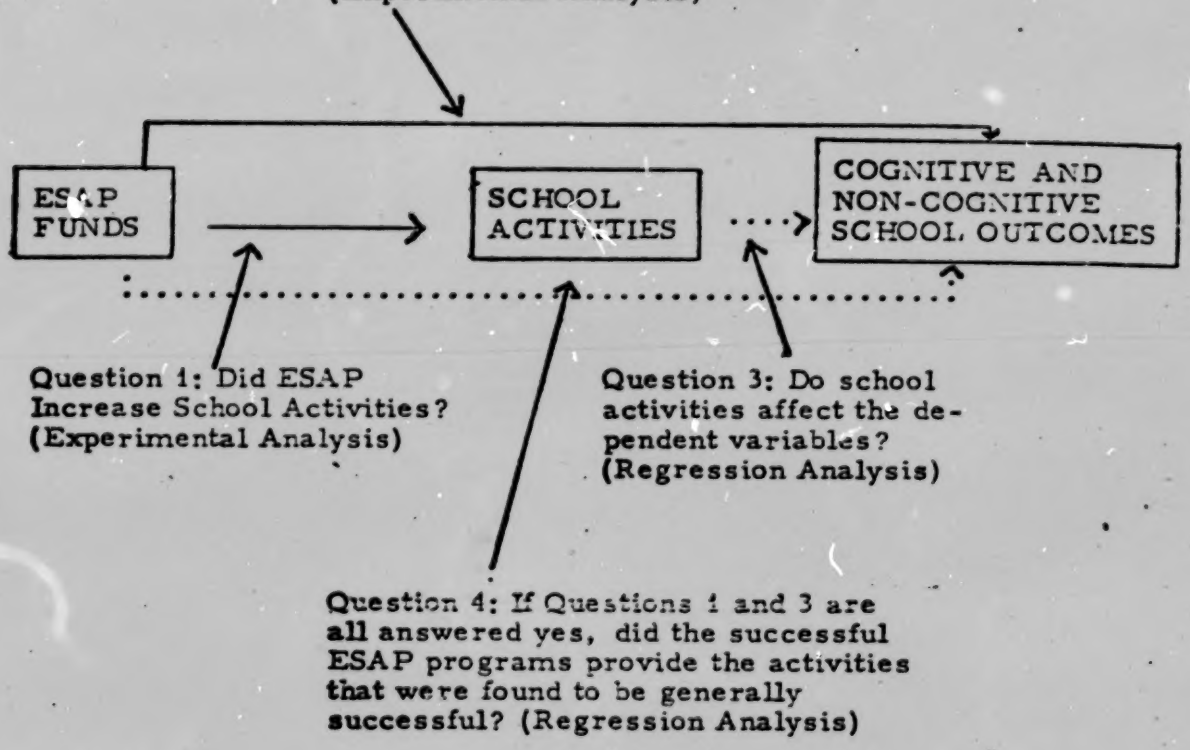
evaluate some programs which had been well-funded and had several years of experience in the school. These programs were, of course, much more likely to be successful than new programs developed with the meager ESAP funds. Thus, even though the ESAP treatment might show no effect, we well might be able to conclude that because ESAP funded (in a modest way) remedial reading, and because remedial reading raised cognitive test scores, then it would follow that a larger ESAP program operating over a longer period of time would be able to raise test scores as well. In short, we assumed that remedial reading regardless of its funding source is essentially the same. When ESAP failed, we would have a chance this way to show not that it was wrong-headed, but only that it was too little and too recent.

(Figure 1 about here)

This straightforward and seemingly intelligent approach to the problem in some ways nearly strangled the ESAP evaluation, as we shall see.

The analysis worked out almost the opposite way. The experimental design did not fail; we found sizeable achievement gains for black male high school students. It was the regression analysis which came close to failure; only after very torturous analysis were we able to locate any school characteristics which seemed to have favorable effects. At the end of the study, we were able to draw three main conclusions:

Figure 1. The Four-Step Evaluation Design


Did ESAP affect the dependent variables? (Experimental Analysis)
ESAP FUNDS
SCHOOL ACTIVITIES
COGNITIVE AND NON-COGNITIVE SCHOOL OUTCOMES
Question 1: Did ESAP Increase School Activities? (Experimental Analysis)
Question 3: Do school activities affect the dependent variables? (Regression Analysis)
Question 4: If Questions 1 and 3 are all answered yes, did the successful ESAP programs provide the activities that were found to be generally successful? (Regression Analysis)

1. ESAP was effective in raising achievement

2. Southern school desegregation is often a special hardship on black male students. Schools where white and black students are present in roughly equal proportions have the highest achievement scores for both white and black students after differences in student socioeconomic status are controlled; in schools where whites predominate by a 2-to-1 or more ratio, all scores are slightly lower, but scores for black male students are sharply lower. This suggests that black males are the important target group for any program to improve quality in newly desegregated schools.

3. By combining the experimental results and a lengthy correlation and regression analysis, we were able to advance a hypothesis about why it was effective: that ESAP funds led to a change in the racial behavior of the school staff, which was observed by black students, causing black male students to like school more, motivating them to perform better on achievement tests.

This set of results led to a specific policy recommendation; that funds should indeed be provided to assist desegregating schools, but that these funds should be targeted toward improving the racial climate of the school, particularly the racial behavior of the school staff.

The evaluation was both a device for making the ESAP program accountable and an aid to future program planning. The knowledge that ESAP was effective is useful data to policy makers, but by the

time we learned this, ESAP had been replaced by The Emergency School Aid Act (ESAA), a program similar to ESAP in many respects but having more of a nationwide scope.[13] Thus, it is fortunate that the evaluation provided an explanation of why ESAP was effective, for this explanation could be used to guide policy planning on other programs.

PART II: THE STUDY

A. DATA COLLECTION

Data was collected from 5th grade and 10th grade students in some 400 elementary schools and 200 high schools.

The following data was collected:

In each school:

Questionnaires and achievement tests filled out by three randomly sampled classes of students (usually three fifth grade classes in elementary schools, or three tenth grade English classes in high schools). An average of 55 students completed the questionnaires in each school.

Questionnaires completed by 10 teachers in each school.

A personal interview with the school principal.

In each school district:

An interview with the administrator in charge of the expenditure of ESAP funds.

Four community leaders (by telephone interview).

Taken together, the questionnaires that were used in each school described the way in which ESAP funds were used, the school's special programs and supplementary personnel, the social background of the students, the quality of race relations, some aspects of the school's "social climate," and the performance, attitudes, and aspirations of the students.

The interview with the school district ESAP director was used to determine how ESAP funds were allocated, both in the district and in each school. The four community leader interviews were designed to give us a description of community factors (such as the level of civil rights activity) that might affect the schools. In almost all cases, the four community leaders were two blacks and

two whites active in the community, but with no professional connection to the school system.

The school principal was asked to describe the programs and staffing of the school, and to give some statistical data, such as the dropout rate and the number of guidance counselors. We also asked attitudinal questions dealing with racial prejudice, perception of the quality of the teachers, and the like.

The ten teachers were selected so as to maximize the possibility that the students studied would have been taught by those teachers. The teachers' questionnaire focused on the teachers' attitudes toward their students, and on their perceptions of the quality of race relations and of the classroom climate. Measures of racial prejudice and attitudes toward teaching in general were also taken.

The questionnaire administered to the students dealt with their perceptions of their school and teachers, and their participation in various remedial programs. They were also asked a series of attitudinal questions related to motivation, happiness, and orientation toward school. After the questionnaire was administered by the interviewer, the students took a one hour achievement test. We used "The Survey Test of Educational Achievement" developed by Darrell F. Bock, consisting of ten to fifteen items selected from each of five subtests of the fourth and ninth grade Educational Testing Services' STEP batteries. The subtests were reading comprehension, language, mathematical concepts, mathematical computation, and science. If we had been interested in individual achievement, the

standard five hour version of the test would, of course, have been preferable. But for our purpose -- the analysis of the mean achievement -- the reliability range of the one hour test (. 84 to . 91) was quite satisfactory.

Having blundered in underestimating the potential of the experimental design, NORC compensated with three wise decisions which turned out to be important. First, the decision was made to aggregate all the data to the school level. Thus instead of reporting the score of an individual student on a particular scale, we instead would report the percentage of students who said "yes" to various questions, or the mean achievement test score for the school. Similarly, we would report the percentage of the ten teachers who gave certain responses. In doing this, the data are transformed from the level at which they are collected -- the individual student and teacher -- to the level we are actually concerned with -- the school. We are interested in measuring the effectiveness of schools; ideally we should have a single measure of the "output" of each school, but there is no way to collect such data except from the individual students who are the schools' "product." Since we were concerned with school effects

rather than individual effects, the aggregation of data focused the analyst's attention on that fraction of the variance in student behavior which could conceivably be explained by school effects. In effect, the within-school variation between students was ignored. Thus what we are reporting here is an analysis of schools. We are not concerned with determining which students had favorable attitudes toward integration, but rather which schools have students who are generally more favorable to it. Our report was fundamentally intended to help make policy and it is these between-school differences which are of interest to policy makers. Working at the school level also minimizes some of the severe problems of response error at the individual level.

Second, the data was aggregated and the analysis carried out separately by each race. Thus, by separating whites and blacks we were able to assume that white and black students would be affected in different ways by school factors. For example, the race of the principal (if it has any effect) should have a stronger effect on students of one race than on the other.

Third, if we thought that students of different races might react differently to school programs, then it also made sense to consider the possibility that boys might react differently than girls. Since it cost relatively little to do the computer work, we stored for future use separate test scores and attitude scores for black boys, black girls, white boys, and white girls.

B. INTO THE FIELD

After a fall of questionnaire writing and two months of waiting for the Office of Management and Budgets' approval, the study went into the field in March of 1972. Here NORC had the chance to show off the quality of its field work. Asking questions about race relations in newly desegregated southern schools is not the friendliest thing for a Yankee research organization to undertake. At the same time, one could assume that the least timidity on the part of the research organization would only encourage politically sensitive school systems to change their mind about cooperating in the evaluation.

The NORC field operation somehow combined the delicacy of brain surgery with the organization and determination of a McGovern primary campaign. There were only a few incidents--a couple of agressive black teachers in a border city, a nosy principal reading supposedly confidential questionnaires somewhere else, an attack from a white anti-busing group in Florida--but a lot of plane trips: senior supervisors were flown in from Seattle to supervise part of the Southeast, and in one case a home office staff person flew 2,000 miles to persuade a superintendent to permit the study, returning the same day. When the field work period ended, data had been collected from students and teachers in all 598 schools and interviews had been conducted with all but two of the principals. With very few exceptions, the interviewers reported that they had been greeted with extreme courtesy and a great deal of cooperation by the local schools.

C. THE REGRESSION ANALYSIS

The process of reducing several hundred bits of data on each of 30,000 students, combining it with the reports of 6,000 teachers, 600 principals, 400 community informants and 100 district administrators was formidable.

Fortunately the teacher and student data was collected on optically scannable instruments which were processed efficiently (by National Computer Systems). The programming to combine the questionnaires, build scales and aggregate the data to the school level took two precious months. The original contract called for the delivery of preliminary results two months after the study came out of the field and a final report four months later. The preliminary report was delivered only a few weeks late, mainly because York volunteered to change the contract to permit an oral presentation. Two oral presentations were made internally to the Office of Planning, Budgeting and Evaluation; the first, only seven days after a completely clean data tape had been created, was interrupted when the 400 pages of computer output on which the presentation was based were found to have a programming error which systematically rendered every single page useless. It turned out that when this analysis was redone (over a weekend!) for the second internal presentation, OPBE got cold feet deciding that the limited findings then available would be of little value in settling any of the issues that the ESAP stuff was then concerned with; the formal preliminary briefing was cancelled.

In retrospect we feel that preliminary reports (written or oral) presenting results to be used to guide policy are undesirable, however valuable they may be for policy-making--in this case, providing guidance about what types of ESAA applications should receive priority for the following year. It must be made clear to program people that the results of preliminary reports are subject to change. Not entirely confident of the preliminary findings, aware of the contradictory evidence provided by different evaluation studies, and mindful of the damaged credibility if the final results contradict the preliminary results, policy-makers could hardly be blamed for looking at preliminary results with a jaundiced eye. And so should evaluators for the same reasons (to say nothing of the inefficiency and extra work prelimianry reports of results create for the researcher).[14]

From the beginning the analysis of the experimental design and the cross sectional regression analysis were segregated from each other. Analysis of covariance (the proper statistical machinery for the analysis of an experiment) is the province of educational psychologists and the average survey researcher is completely inexperienced with this method. A graduate student psychometrician--Carlyle Maw--joined the project to carry out the analysis of covariance. Meanwhile, the rest of the staff pursued the multiple regression technique they were more accustomed to, firmly convinced that they were going to "save" ESAP from the experimental design. (Recall the unflagging conviction of NCRC that the prospects

of ESAP showing a measurable gain in any student outcome were nil.) The regression analysis began with an attempt to explain achievement test scores by looking at the impact of various compensatory education programs and other school characteristics ranging from the presence of gym teachers to the number of textbooks purchased in the last two years.

It seems fair to say that 95% of this analysis was boring to the point of disaster. Something less than 1000 regression equations were computed. The regression model involved examining the data to find the most important control variables--the predictors of achievement which could be considered logically prior to ESAP programs (mainly student background characteristics). This produced separate control equations for each grade and racial subgroup. The second step involved placing the control variables in a series of multiple regression equations along with one program or activity as the independent variable and mean achievement as the dependent variable. In brief, the 60 program or activity variables were based on three sets of questions: special personnel (not including regular classroom teachers), programs (such as tutoring or student relations programs), and equipment or supplies. These equations were repeated until all possible activities had been tested for each of the four grade and race combinations.[15]

These variables included such things as number of remedial reading teachers per capita, presence of a student human relations program, use of student tutoring, and number of teacher's aides. The most frequent result was a standardized regression coefficient (indicating the size of the impact of the program on achievement) of .00. After a while we began to view as "positive" coefficients as small as +.06.

Part of the reason for our trouble is that school programs which a school administrator might consider quite worthwhile have really quite small impacts when viewed in the overall scheme of things. Consider the following example. Suppose an elementary school embarked on a particular program which had a cumulative effect by 5th grade of elevating the achievement test scores of a group of students by approximately one-half a grade. While this is not an awesome effect, a school official who believed that test scores were valid criteria of effectiveness would certainly judge this program to have been successful. Let us further assume that we are able to rank schools from those where this program is completely absent to those where it is present to a modest degree up to these which have a full-blown program. After removing (as best we can) the effects of student background characteristics,[16] a plot of school mean test scores against presence of the program might look like that of Figure 2. But the size of the effect in Figure 2 is not very large. The standardized regression coefficient is given by the formula

$$B = \frac{\text{vertical gain}}{\text{horizontal distance}} \cdot \frac{\text{standard deviation, horizontal variable}}{\text{standard deviation, vertical variable}}$$

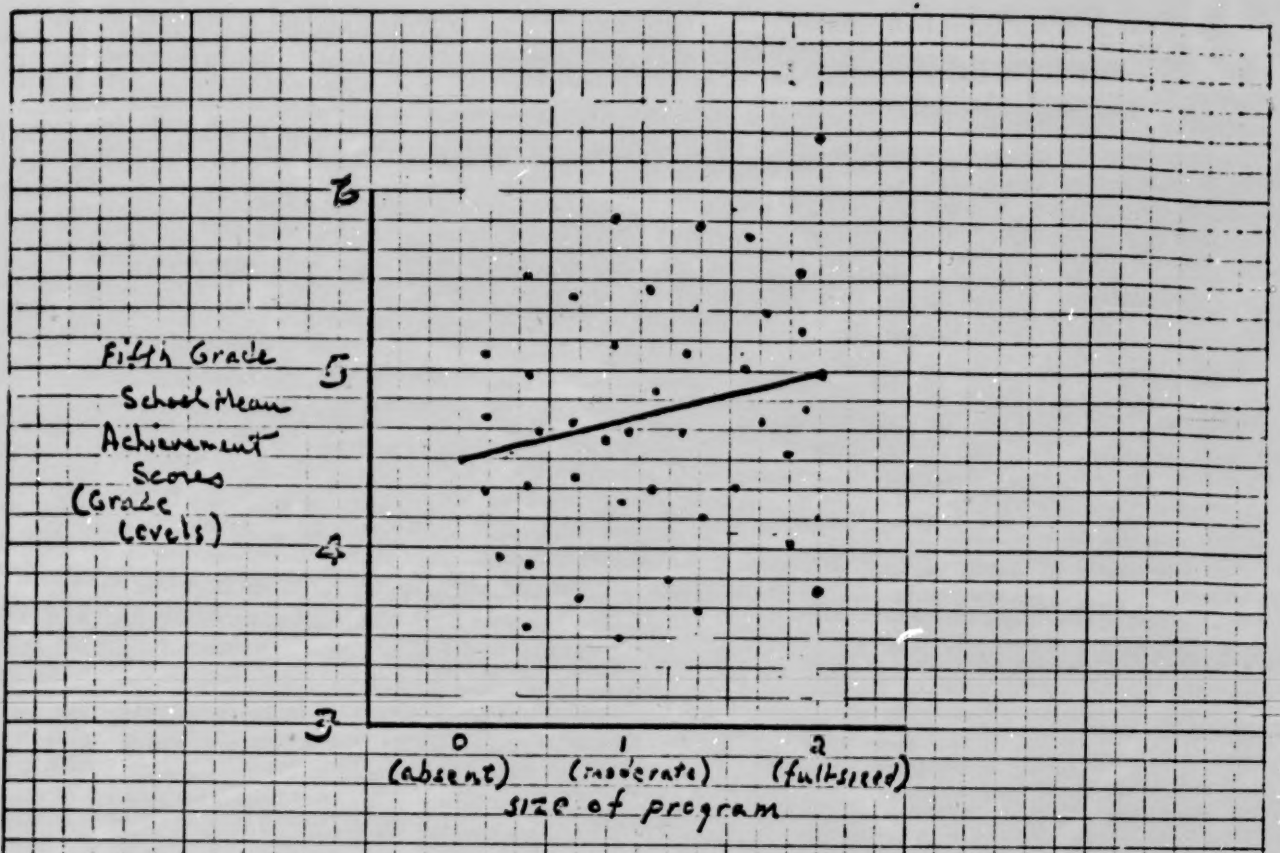


Fig. 2: THE EFFECT OF A HYPOTHETICAL SCHOOL PROGRAM ON ACHIEVEMENT

For these data we might assume that the standard deviation of the independent variable (the program) is approximately .25, and the standard deviation of the mean school achievement test scores (after the effects of social effects are removed) is approximately one grade level. Thus, $B = \frac{.5}{1.0} \times \frac{.25}{1.0} = .12$[17] With 400 cases the effect would barely reach significance and the reader would be unimpressed by a coefficient of "only" .12. The problem is that the effect of one-half grade level has been restated into a different scale and seems to have become a small number (.12) in the process; to make it worse it can be restated as explaining less than 1% of the variance in achievement. Of course, the effect is small, compared to the total effect of student socioeconomic status, but it is not necessarily small compared to other known alternative methods of education. We are justifiably cautious in interpreting regression results because we know that it is affected by error of measurement. Generally, however, the error of measurement causes the regression result to itself be conservative. For example, in measuring social background measure for black elementary school students is only .35 -- meaning that if we had a perfect measure of social status, it would correlate with our measure only $\sqrt{.35}$ or .6. We can get rid of some of this error by computing the average social status for each school. When we do this, some of the individual errors cancel each other out, but the total reliability of the average social status of the black 5th graders in a school rises only to .53 -- the correlation with the "true score" being $\sqrt{.53}$ or .73.

The effect of this error of measurement is to make our regression analysis understate the effects of school characteristics. For example, when we attempted to assess the impact of a school having a remedial reading teacher on Black 5th grade test scores, first, we found that the simple correlation between the presence of a remedial reading teacher and the achievement level of students was -.15, indicating a very definite tendency for remedial reading teachers to be assigned to those schools with the lowest Black achievement scores. When we carried out the multiple regression analysis in order to control for effect of student background the standardized regression coefficient became a negligible +.01. But suppose we believe for a moment that our socio-economic status variable does indeed have a reliability of only .53. Then the standard recalculation of the data (known as a correction for attenuation[18]) indicates that had we been able to measure student background factors perfectly, we would have found the apparent impact of remedial reading teacher to be +.19--indicating that student test scores were nearly a grade level above expected when a remedial reading teacher was present. Thus, despite out suspicion that remedial reading was having a decisive effect, we nevertheless were forced to draw a conservative conclusion--it is difficult to argue that a regression coefficient of +.01 really means a strong positive effect.

In summary, one problem with multiple regression is that it tends to produce results--regression coefficients and percentage of variance explained--in units which are misleading, tempting even the trained reader to judge that effects are small. But second, multiple regression simply doesn't work when there is measurement error and when the two groups being compared are very different from each other, as was the case with the remedial reading analysis.

After assessing the impact of dozens of different school resources, we were able to draw only one convincing conclusion. We found achievement test scores for both black and white students to be markedly higher in the small group of high schools which claimed to have an audio-visual specialists. Vanecko thought the result convincing and argued that it was not the utility of audio-visual equipment for teaching, but its impact in moderating racial tensions that caused the achievement gains. York argued that too many other evaluations of audio-visual equipment usage had found no effect. We finally decided to gather more data by contacting the 17 high schools which had claimed to have audio-visual specialists. We found that of the 17, those 12 which indeed had a highly developed media program did indeed have quite high test scores and furthermore they had unusually low levels of racial tension (statistically significant at the .05 level). On the basis of this follow up we were able to recommend that the government make some sort of investment in audio-visual use, at least to the point of further research.

But this one finding was the bright spot in a months-long analysis by Vanecko of possible program effects on test scores. All the other effects stubbornly clung close to .00, and it was beginning to look like our data was supporting the worse fears of the "schools can't make a difference" viewpoint.

D. SCHOOL EFFECTS ON RACIAL ATTITUDES: A FACTOR ANALYSIS

Fortunately, we did test for program effects in one non-cognitive area: the attitudes of students toward desegregation. Here the story was much more interesting. By factor analyzing the same school characteristics used in the search for achievement effects, we were able to isolate what seemed to be three alternative approaches to education. Some schools emphasized cognitive development more or less exclusively. A second group of schools had more highly developed programs built on a therapy model, emphasizing guidance, counseling and intensive use of social work professionals. A third group of schools stressed reform of the curriculum, teacher in-service education, and a strong emphasis on human relations. When these three factors were entered as independent variables in the regression analysis of racial attitudes, the third group of schools had consistently more favorable attitudes toward integration on the part of both black and white students. This suggests that those schools which were committed to good human relations, and which recognized the need to change the attitudes not only of students but of staff as well, were more successful. This would seem to suggest that the new reforms in elementary school education, built around individual instruction, open classrooms, etc. may halp. It also suggested that the basic ESAP strategy of using federal funded projects to improve race relations, might be workable. Even here, however, our data were not clear enough on this point to serve as convincing evidence, although this tentative conclusion would prove useful in analyzing the experimental design.

Thus, the regression analysis brought us some interesting results, but at the same time left us with the feeling that we were either mining low-grade ore or else using very dull tools.

E. THE EXPERIMENTAL RESULTS

Meanwhile, the analysis of the experimental design was as simple as the regression analysis was complex. Because both white and black scores were involved, Maw elected to use a multivariate analysis of covariance, a method of comparing the mean achievement of the experimental and control schools which could not only look at effects on each race separately and take account of possible differences in social background characteristics between the experimental and control schools, but which could also combine the white and black scores to produce a single test of signficiance. Since a multivariate analysis was necessary, and since separate scores for girls and boys were available in the data tape, it was natural of him to analyze the results by sex. To the best of our recollection no conscious decision was made at that point in the study to look at sex differences; it was something that just happened.

The results showed that the black male achievement in both elementary schools and high schools was somewhat higher in the experimental schools. Furthermore, when the effects of social background were taken into consideration the difference for the high school students became statistically significant ($p < .02$). Obviously, the more categories we divided the students into (grade, race, and sex) the more likely we were to get one significant difference; hence we needed the multivariate test, which found a

significant effect in high schools when the results for males, females, whites and blacks were considered simultaneously. The results are shown in Table 1.

(Table 1 about here)

The results seem to be not only statistically but also socially significant. The gain in 10th grade black male achievement indicated by the analysis of covariance, 24 point, is approximately equal to a half year advance.[19]

F. MAKING SENSE OF THE ESAP EFFECT

The results of the experiment may answer the question about ESAP's effectiveness but they raise a host of others about why this program should be effective. Bear in mind that the experiment does one thing and does it well -- it can tell us that the schools which received the treatment performed better than those which did not. But the experiment cannot tell us anything more than this; it cannot tell us what the treatment really was, or why it worked. These two questions are especially important in the case of ESAP. Given that the total amount of money was quite small, and the duration of the program quite short, it is hard to imagine what ESAP did that was so effective. Thus, finding a positive ESAP effect led us into the more difficult task of deciding how ESAP worked.

The first step was to compare the ESAP results to the regression analysis. Unfortunately the regression analysis had been done with sexes combined before we got the experimental results, and we lack the will to redo it. We did however construct an overall scale of school

Table 1. ESAP's Effect on Achievement: The Results of the Experiment Design

Grade, Race, and Sex	Unadjusted Scores			Difference: Experimental-Control	Difference Adjusted for Social Background	Significance Level[a]
	Standard Deviation	Mean Experimental	Mean Control			
Fifth grade white male . . .	46.8	322	318	4.2	15.3	n.s.
Fifth grade white female . .	42.1	358	362	- 4.4	- 5.9	n.s.
Fifth grade black male . . .	63.3	160	146	14.2	- 3.8	n.s.
Fifth grade black female . .	55.4	193	192	1.0	- 1.6	n.s.
Tenth grade white male . . .	63.5	252	241	10.4	[illegible]5.7	n.s.
Tenth grade white female . .	52.6	276	273	3.2	-15.5	n.s
Tenth grade black male . . .	38.4	117	102	14.9	24.0	$p < .02$
Tenth grade black female . .	43.5	123	120	2.9	- 4.7	n.s.

NOTE: Multivariate significance test using all four dependent variables combined in a linear model:
Fifth grade: n.s.
Tenth grade: $p < .04$.

projects, weighting each project according to the likelihood of ESAP funding it; this scale, a sort of "ESAP components scale" was positively related to black male achievement (B = .08), although not as strongly as ESAP itself was. This led to our first conclusion: that ESAP was, on the one hand, effective because of the activities it funds, but on the other hand, ESAP was also more than the sum of its parts.

In many evaluations one can understand the results in terms of reasonably tight economic models or simple input/output analyses. But here we had to build virtually a theory of education in order to understand what was happening. There were a number of findings in the study which began to come together to explain the ESAP effect.

1. In a number of cases our data indicated that factors of motivation and school morale were of great importance in explaining differences in achievement test scores. For example, one finding (which was <u>not</u> included in the ESAP report) was that schools whose football and basketball teams had winning seasons also tended to have markedly higher achievement test scores.

2. The analysis of racial attitudes had shown us that different school policies resulted in significant differences in student attitudes toward race.

3. A wide variety of findings indicated that black students were extremely sensitive to the racial climate of the school; for example the single best predictor of the degree to which black students described themselves as "happy" is the percentage of black students who felt that their teacher and principal were in favor of integration.

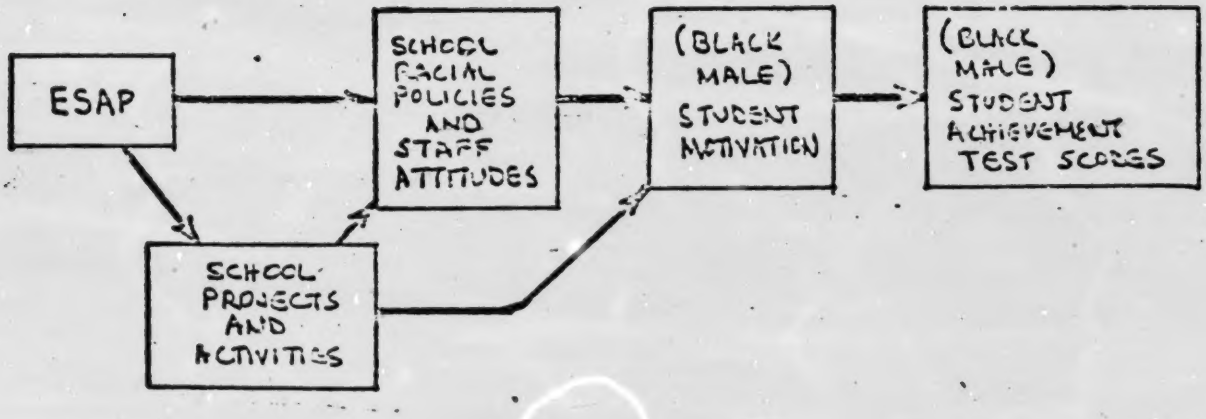
4. Finally, other analysis argued that it was logical that racial factors would affect motivation, and hence the test scores, of black male students. Two Johns Hopkins graduate students, Janet Griffith and Ruth Narot, analyzed the effects of racial composition of the school on achievement test scores. In general, they found that in desegregated schools, as the number of white students increased, black achievement rose, but that if the school was over 70% white, the achievement of black male students dropped sharply, while black female achievement remained high. Apparently, something about the racial climate of predominantly white schools was preventing black male students from achieving. Thus it made sense to argue that ESAP's main effect was in reducing whatever racial tensions were disrupting black male achievement.

In summary, these pieces of data supported the model of school achievement shown in Figure 3. First we had shown that in general motivation has an impact on achievement, then we showed that for black students, and especially for black male students, racial characteristics of the school were strongly related to his motivation. Third, we had seen that the racial climate of the school can be affected both by the school policies and by the behavior of its teachers. If we take the right hand portion of the figure as being supported by these data, our question becomes "How did a small program like ESAP affect either the organizational structure of the school or the behavior of teachers so as to produce a change in racial climate?" Figure 3 hypothesizes two possibilities: that ESAP affected the school by changing its mix of projects and activities, or that the simple presence of ESAP, the knowledge that funds had been provided to make desegregation work,

had a direct effect on the staff.

At this point Ruth Narot undertook an analysis of teacher racial behavior to determine if teacher behavior should be viewed as a relatively unchangeable character trait, or whether informal pressures could be effective in changing teacher behavior.[20] Her analysis indicates that behavior can be changed; for while the personal feelings of teachers about racial issues such as open housing or intermarriage seem related to background characteristics such as age, sex, or place of birth, both the teacher's feelings about school integration and the perceptions of her feelings (either by her staff or by students) were not simply a matter of personal background; they were strongly affected by the racial attitudes of their principal, the amount of civil rights activity in the community, and the degree to which the principal attempted to influence his teachers to support desegregation. Narot's analysis added another stone to the argument we were erecting. The fact that the presence of civil rights activity in the community is associated with teachers behaving more liberally in public than they feel in private suggests the degree to which social forces, either appealing to the conscience of whites or attempting to coerce them, can result in widespread behavior change. Showing that teacher's behavior can be changed in this way makes it more plausible that ESAP could have created pressure on teachers to change.

Figure 3. A Theory of ESAP's Impact


ESAP
SCHOOL RACIAL POLICIES AND STAFF ATTITUDES
SCHOOL PROJECTS AND ACTIVITIES
(BLACK MALE) STUDENT MOTIVATION
(BLACK MALE) STUDENT ACHIEVEMENT TEST SCORES

G. HOW ESAP FUNDS WERE SPENT

With this conceptual scheme developed we were prepared to believe that ESAP's impact on achievement might have been because of the way in which it affected school race relations. We next returned to the experiment and began searching through the school characteristics and school programs to see where ESAP had an impact. Here we discovered the other great benefit of the experimental design. A chronic problem in most evaluation research involving federal funds spent by local administrative units is tracing the funds to find out what precisely was done with them. But with an experimental design we know that any difference between the experimental schools and the control schools (within the known probabilities of sampling error) was the result of ESAP. We had simply to list the programs available in all the schools and compare the two groups. When we did this, we found the biggest difference in elementary schools to be the presence of remedial reading programs. Sixty-two percent of the experimental schools had created a remedial reading program compared to only 46% of the control schools. The experimental elementary schools were also more likely to have counseling programs and teacher's aids. In terms of the three alternative strategies for improving education defined by our factor analysis of school programs, it would seem that most of ESAP's funds were devoted to traditional cognitive development programs.

The pattern of expenditures is what one would expect from a small program with short-term funding receiving during the school

year. New professional staff could not be employed because of lack of time. Similarly, major educational reforms could not be undertaken without more stable, long-term sources of funding. Consequently, all the ESAP effects on school programs were the result of either the use of sub-professionals, the purchase of additional supplies, or payment of stipends to teachers to attend special workshops.

The high school pattern was more complex and less traditional. The most important single difference between the experimental and control high schools was in the percentage of schools having "human relations programs designed for teachers". Sixty-four percent of the experimental schools had such activities according to their principals, compared to only 40% of the control schools. The next large difference was in the number of high schools saying they had revised their curriculum this year: 70% of the experimental schools compared to only 47% of the controls. Other differences were an increase in extracurricular activities geared toward minority students, and an increased amount of work with parents. Other smaller differences were consistent with the general emphasis of ESAP funds upon intergroup relations problems.

If we ask why ESAP funds were used differently in high schools and elementary schools, the most compelling answer is that high school students are less passive than elementary school students. Being less passive, they are not only likely to create a more unpleasant racial situation, but are also likely to make sure that the school administration

pay attention to it. If elementary school teachers and principals believed that race relations was a problem, they could still reason that the solution to the problem is to change the students, through traditional remedial programs. In high schools, it seems likely that school administrators had to admit that race relations were a problem, and that their chances of changing their students were slim. Therefore, we argue that the elementary schools considered ESAP an opportunity to do what they prefer to do -- basic instruction-- while the high schools saw it as an opportunity to do what they had to do-- change their curriculum and their staff's behavior.

H. THE EFFECT OF ESAP ON SCHOOL CLIMATE

The final task of the analysis was to determine whether the experimental high schools did have a better racial climate, and whether this was related to the improved achievement test scores of black males. We examined all of the measures describing the school and found 10 measures where the difference between the experimental school and the control school was .2 standard deviations or more. Five of these were related to racial issues and they are shown in Table 2. The table shows that in ESAP high school teachers were more likely to discuss racial issues and less likely to see the school as racially tense, while the black students were more likely to perceive the school staff as supporting integration, more likely to say they liked school, and less likely to say "I feel like I don't belong" in this school. In fact, the experimental school staff did not appear more liberal in reporting their private attitudes toward race. Thus, ESAP seems to

have changed either the way these teachers acted (not the way they felt) or the way their actions were perceived.

Table 2. Tenth Grade Experimental and Control Schools Differences in Racial Items

Item	Difference, Experimental School Minus Control, in Standard Deviations
Teachers: discussion of racial issues more than once per month	.30σ
Teachers: school not racially tense	.35σ
Black students report staff is pro-integration	.40σ
Black students: "I feel like I don't belong" (per cent yes)	-.24σ
Black students: "I like school" (per cent yes)	.26σ

It appears that ESAP schools have a more egalitarian racial atmosphere than their matched controls, and are more likely to be places where black students like school and feel a sense of belonging. All of this is important in itself. Very few would question the value of making black students feel less alienated. But since this is an analysis of achievement, we need to explore the relationship between black students feeling more welcome and their academic performance in school.

Of all the items in Table 2, the item showing the strongest relationship with achievement is the percentage of students who say that they like school. Taking the tenth grade sample of matched pairs, we computed the correlation between (1) the difference between the experimental and control schools in the percent of black students

who say they like school and (2) the difference between the experimental and control schools in black male achievement. The correlation between the two differences is . 50. In thos e pairs where the ESAP school had a higher percentage of black students who like school, black achievement was also higher.

Since this is a correlation between schools, and is much stronger than the correlation among individuals (at the individual level, r for black males is . 15), we cannot argue that this is simply because students who do well academically like school. Moreover, there is evidence that for blacks, liking school has a definite racial meaning, and one that is quite different from its meaning for whites, as Table 3 shows. (Table 3 about here)

The percent of students liking school and feeling as though they "belong" are highly correlated for both races (. 37 for whites and . 30 for blacks). When we look at perceptions of staff attitudes, however, we see a sharp white-black difference (Table 3). The table shows that perceptions of the racial attitudes of the staff are highly correlated with black sense of belonging and liking school, and are not at all important for whites. We think the most plausible interpretation of Table 3 is that as black students perceive that staff attitudes are more pro-integration, they feel less alienated in their schools and find it easier to learn.

Presented this way the argument may sound persuasive; actually it represents merely the most convincing discussion we could put together in several months of struggling with the data. The problem is, of course, that the experimental design was intended to evaluate

Table 3. Zero-Order Correlations Between Liking School, Belonging, and Perceiving Staff as Pro-Integration, for Tenth Grade Black and White Student Bodies.

Item	Per Cent Who Say They Like School	Per Cent Who Say "No" to "I don't belong"	Per Cent Who Feel Staff is Pro-Integration
A. Blacks			
Per cent who say they like school	-	+.30	+.44
Per cent who say "yes" to "I don't belong"		-	+.44
Per cent who feel staff is pro-integration			-
B. Whites			
Per cent who say they like school	-	+.37	+.09
Per cent who say "yes" to "I don't belong"		-	-.07
Per cent who feel staff is pro-integration			-

ESAP. While it did its job very well it did not and could not tell us what ESAP "really" was. The experimental design also could not help us in determining which particular ESAP programs were effective. As soon as we began comparing one experimental school to another, we no longer had an experimental-control situation, but had to fall back upon standard cross-sectional statistical techniques (such as multiple regression, of which we have said too much already). If we found that those schools which spent their ESAP funds on staff in-service education tended to have larger achievement differences from their control schools than those experimental schools which spent their ESAP funds on remedial reading, we would still not be able to conclude that one program was effective and the other ineffective; perhaps teacher in-service education is a program which is adopted only by schools with high achieving students. In the experimental design, there is no question about why an experimental school is an experimental school rather than a control; it is one or the other because we made it so.

There remains one final mystery about ESAP, one which our data cannot unravel. If these considerable achievement gains could be obtained merely by exerting pressure on teachers to adopt more liberal attitudes, why could not the school principal have done this without ESAP funds? What could he do with $10,000 that he could not do without it? One argument is that $10,000 is actually a great deal of money when we consider how small a fraction of the schools budget is available for the discretionary

use of the principal. If he wishes to schedule a one-day workshop in which to encourage his teachers to work on racial matters, he may find there are no funds in the district budget to pay expenses or salaries to them. If he wishes a new multi-ethnic text, he will need to pay for it with something.

It is also possible that ESAP funds were not the important matter; ESAP itself as a symbolic gesture is what made the difference. The principal who received ESAP funds knew that he had received federal aid to help make desegregation work. Having received these funds, it was therefore reasonable for him to set about trying to both spend the funds and made desegregation work. What ESAP may have provided then was an instruction -- attempt to improve race relations! --and a legitimation. You are expected because of having these federal funds to work to improve race relations. At first, such a hypothesis seems unlikely. After all, the principal surely must know that race relations is a problem and certainly the principal has the authority to direct his teachers in their work. But let us pursue the questions further. Under what conditions does the principal have the moral right to ask his teachers to treat black and white students in particular ways? Isn't it just as reasonable to argue that the wise principal will recognize that teachers will have personal views and be respectful to them? Is asking a conservative Mississippi school teacher to teach black history any different

than asking a Catholic teacher to teach birth control?[21] But the presence of ESAP settles this issue; the principal is mandated.

I. THE IMPACT OF THE EVALUATION

As we write this, the report is only a few months old. A few friendly responses encourage us to believe that the report may have some effect, but it is too early to know. The program itself has been in a constant state of change; the Emergency School Assistance Program was a temporary program that expired when Congress passed the Emergency School Assistance Act (ESAA) in 1972. ESAA is of course being evaluated now. Since the major component of ESAA--School District Basic Grants--is more similar to ESAP than different, we may hope that the evaluation has some impact on it. Already a quiet effort has been made to alter the funding priorities of ESAA to encourage more human relations effort.

The ESAP evaluation may be used as amunition by the non-cognitive faction within education in their perennial battle with the cognitive development school of thought. It is unlikely that any single evaluation will have great impact upon such a serious issue and indeed it would be unfortunate if an important shift in thinking on these matters were a matter of a single evaluation report.

The Office of Education has instituted a formal mechanism to attempt to assure that evaluation studies will have an impact on policy The procedure is that the project officer for the evaluation in the Office of Planning, Budgeting and Evaluation (in this case, York) drafts a Policy Implications Memorandum (PIM) to the Commissioner of Education. This Memorandum presents discussion and specific recommendations in the areas of legislation; budget; management; and planning, evaluation and research which can be drawn from the evaluation. The recommendations for action are to be as specific as possible as to what is to be done, by whom and when. Recommendations

should be realistic and practical of accomplishment and take into account current and proposed administration policy and practices, availability of resources, political realities, etc. When the PIM is drafted every attempt is made to negotiate agreement with the affected Deputy Commissioner[22] and program staff prior to obtaining the Commissioner's approval. When agreement cannot be negotiated, a cover memo to the Commissioner highlights the disagreement and requests a decision.

The PIM process follows delivery of the final report, preparation of an executive summary, and release of the report to Congress. Thus, the PIM process is not terribly speedy. In this particular evaluation, two briefings by Crain and one by York preceeded the PIM. Involving the Deputy Commissioner and program staff and legislation staff, the briefings led to the decisions on emphasis on human relations programs mentioned earlier in this section. This was necessary because this decision regarding funding for second year ESAA grants was needed in early Fall. The PIM for this study, containing other recommendations, is still in draft stage at this writing.

The ESAP report appeared at a time when many persons were

beginning to express sympathy for experimental design in evaluation research. Thus, we could expect it to be cited frequently by proponents of the experimental methods. We can only hope that if it is a harbinger of other similar research projects, but that they learn from our mistakes as well as our successes.

J. A MORAL

One of the reasons why we struggled so hard to find a convincing explanation of why ESAP was effective is that we knew that our audience would be convinced before the fact that ESAP would have no effect--just as we were convinced at the beginning of the evaluation. Had we held open even a forlorn hope that the experiment would show a positive effect, we would have planned our analysis differently and probably arrived at a clearer picture of what ESAP did in considerably less time. Perhaps instead of asking "Why did ESAP work when sociological common sense tells us it could not have" a better question would be to ask "Why does common sense tell us that a federal program designed to improve schools must necessarily and inevitably fail to accomplish its mission?"

The answer is that it is not common sense that tells us. Common sense does not tell us that a dollar spent on education will accomplish nothing anymore than it tells us that 20 cents put on a lunch counter will never produce a cup of coffee. Rather is it the "uncommon sense" of intellectuals and professional evaluators. If there was a shared ideology among the members of the intellectual

left in the late 1960's (and this included the vast majority of social scientists) it is that authority is evil and institutions incompetent.

Then, when the Coleman Report showed that "only" 20% of the variance in cognitive test scores lay between schools, the intellectual community jumped to agree that indeed the differences between one school and another were of no importance. Indeed, very few sociologists questioned this interpretation of the data (although Coleman himself has done so recently).

The other ideological strain runs through the thinking of professional evaluators, who struggle to protect their professional integrity by proving over and over again that they are not bought by the government whose work they evaluate. The radical criticism of social science is that it is the paid servant of a conservative government, biasing its work to "blame the victim" for his poverty rather than social institutions. A good case can be made for this view. But it seems to us the opposite criticism -- that evaluation research is biased toward criticizing the social order, in order to demonstrate its independence and its intellectual superiority--is equally true.

K. CONCLUSIONS

This has been a case study of the evaluation process. The story is of a success. Living through this project also makes one realize why most RFP's require proposals to contain a section called "corporate capability." It would have been impossible for a university researcher to do this evaluation without the support of a

seasoned and capable organization. But this case study also illuminates a number of the problems that seem to consistently plague evaluations.

This project is unusual because the senior staff held full-time academic appointments and NORC was a more-or-less academic institute. The difficulties that NORC had pointed out why so few academic research institutes do contract research. NORC, lacking a large resident research staff, was dependent on two non-resident sociologists and we found that regulating the deadline of a research contract around the early morning fog at Chicago's O'Hare airport was a serious problem.[23] Another reason why academic scientists and University based research institutes are so rarely involved in contract research other than as occasional consultants or members of an Advisory Committee, is that university researchers have the resources to insist on working in the narrow specialities of their interest. This means that the research institute may find itself with several highly-paid researchers who give the institute a capability only to work within very limited research areas. This in turn means considerable risk of financial problems.

Do academic research groups do better evaluations than profit-making organizations? Probably not, although in this particular case an academic group may have been a wise choice

for one reason: the decision to analyze the data separately by sex. This is the sort of creative (or merely clever), professional (or merely irresponsible?) decision which academics are likely to make. Except for this one decision, we would rate academic and non-academic contractors about even, since the academic project makes up in technical ability and professional standards for its lack of responsibility to the client, and its tendency toward sloppy administration.

The issue of comparing academic to for-profit groups is complicated by the fact that NORC is neither fish now fowl--its great strength in questionnaire construction and fieldwork derives from the large amount of contract work it has done, although perhaps the quality of its staff and its professional standards may reflect its academic roots.

One way in which York suffered because of his decision to use an academic evaluator was in the lateness of the report; the deadline for the completed report was six months after completion of field work; this deadline "slipped" an additional eleven months.

Government agencies respect the talents of academics but are ambivalent about employing university researchers for contract research--quite simply because university researchers are usually affiliated with small research institutes with very limited data collection capability, have large amounts of time committed to other research or teaching, and are not dependent on government

contracts for their livelihood and can afford to be prima donnas. University people are unaccustomed to having their work supervised in the way that a research contract is. Thus, it is not surprising that Crain and Vanecko occasionally found themselves furious at York, and vice versa.

There is one final point worth making about the conflict between academic and nonacademic researchers. Despite the obvious way in which the ESAP evaluation could contribute to the knowledge in several areas of basic sociological research (not only sociology of education and race relations, but also socialization and formal organization theory) this type of research is treated as undignified by academics. For example, one member of the research team was subject to evaluation by his academic colleagues in the course of the project; it became necessary to have prestigous academics vouch for the academic respectability of the evaluation, in a way that probably would not have been necessary for a traditional academic research project.

The conflicts created for academics by this distain held for "contract research" do little to help the government, and in the long run are only emabrrassing to the academy.

It has been something of a surprise and a disappointment that there has been very little interest in further analysis of the data. This is the largest data base on school desegregation programs and processes since the Coleman data were collected a decade ago and received reasonably widespread

news coverage.[24] The volume and range of data collected are sufficient to interest anyone from the narrowest policy researcher to the narrowest discipline researcher.[25] The analyses conducted thus far are a small fraction of the possible useful analyses. We suggest a few reasons for this nonresponse: the inadequacy of the news media as a way of stimulating researchers, the absence of any published articles (until now) to supplement a limited printing of a final report, and the desire of researchers to collect and analyze their own data (with its limitations, including small sample size) rather than analyze someone elses (with its own shortcomings). In fact, both the report and the data types with thorough documentation are available.[26]

This case also points out why the separate evaluation group within an agency like the Office of Education is valuable. York was reasonably well insulated from the pressures of the program being evaluated; he certainly had no vested interest in ESAP nor any need to be more than normally reasonable toward the ESAP staff. This

puts him and his colleagues in a position to work effectively to protect evaluations from all sorts of potential bureaucratic interferences.

There is a firm rule in research: "never do research for a client which does not understand research." There is a corrolary: the more the monitor knows about research, the better the product. An evaluation of this size is not a one-man show. York, Crain, Vancko and the senior staff at NORC shared many decisions. While this situation was not always comfortable, it did produce better research.

But the most important lesson from this project is that those introductory lectures we all received on the necessity of objectivity in science have to be taken very seriously. We are accustomed to supposedly scientific work which takes a left or right bias. Being aware of this problem, we were able to work self consciously to avoid interjecting a pro-integration or anti-integration bias in the research. What we were less aware of is the much more widespread bias in evaluation research of disbelief that education programs can succeed. Here we were rescued by the experimental design, which left very little room for misinterpretation. As NORC's Carol Stocking put it: "the trouble with an experimental design is that when you get a result, you have to believe it whether you want to or not."

APPENDIX

A. TWO TECHNICAL NOTES ON EXPERIMENTAL DATA

There are two not-so-minor points with experimental design that the reader may be curious about. First, we noted the great difficulty we had in deciding what it was that ESAP had in fact done that caused the effects which we were able to attribute to it. In our report to the Office of Education, we pointed out this problem and recommended that future evaluation research be done slightly differently.

Secondly, the observant reader may be puzzled by the way in which the analysis of covariance caused the mean difference between the experimental school and the control school to increase. Randomization should not produce this pattern, and we wish to discuss briefly what we think went on here.

B. A SUGGESTED IMPROVEMENT ON EXPERIMENTAL DESIGN

The evaluation was conceived of as an evaluation of ESAP. Such an evaluation makes sense when the policy maker wishes to know whether a particular program should be discontinued or expanded. Frequently, however, the policy maker wishes to know more than this. He wishes to know what kinds of activities result in improvement in order to incorporate these programs into a wide variety of future legislation. In short, what is sometimes called "formative evaluation" is needed. To know that ESAP succeeded tells us only that if 1971 ever comes around again, we should recruit Herman Goldberg and his staff and whip out of mothballs the ESAP program. But in order to know what sort of program to institute

in 1975, when the desegregation issue is different, when public attitudes have changed, when school administrators will react differently to federal guidelines, requires that we know why ESAP worked, what portion of ESAP worked, and what environmental factors interacted with ESAP. This meant that it became critically important, not to show that ESAP worked, but to create a theory to explain why it worked. Since the experimental design could not answer this question, we had to struggle at length with regression and correlation analysis. Regression was helpful--the isolation of the positive effects of human relations programs was the key to understanding ESAP--but the study would have been much stronger had we been able to demonstrate the effectiveness of human relations activities through an experiment.

For this reason we recommended that future evaluations be designed so as to evaluate specific known treatments. Rather than evaluating an entire program, particular activities would be isolated and treated by experimental design.

Consider a simple example. Suppose that we had believed that ESAP could be effective either because it improved race relations or because it improved remedial reading. It would then be possible to ask two southern superintendents to develop these two alternative programs in such a fashion that they could be used by other school districts. Once these two intervention strategies have been developed a randomly selected list of school districts could have been approached and offered ESAP funds for newly desegregating schools in their district

on condition that they adopt whichever one of these programs is designated for that district. This could then be done in a standard "square" experimental design as shown in Figure 4. There would be three treatment groups as opposed to one and the effects of each program in isolation and the effects of both programs operating in tandem could be evaluated. While the point is an obvious one, it represents a marked redirection in the meaning of evaluation research--pushing it more toward a research and development strategy of formative evaluation, and away from the cost-benefit model. In the past evaluation has been seen mainly as an accountability tool--a mechanism by which Congress may determine whether it has spent money wisely. But in fact, this is a very limited use of evaluation and it would be easy to take the machinery of social research and put it to work helping Congress to develop new programs rather than merely determining whether past money was spent wastefully or wisely.

This recommendation was not included by York in the Policy Implications Memorandum because he believed there was too little evidence of specific effective educational activities to justify this recommendation. The major candidate activity was human relations activities but too little was known about specific components of successful human relations activities. Rather, a longer range strategy was adopted. A contract had been signed by this time with Educational Testing Service to conduct a study of candidate exemplary

desegregated schools. The RFP specified a study design including repeated survey data collection plus site visits. As the ESAP results became clearer, ETS agreed to devote much of the site visits to topics that appeared promising as a result of the ESAP study. Thus, the site visits included, among other things, an edamination of human relations activities being conducted in the schools in the study. The intent was thus to provide both additional evidence and detail about effective human relations activities before proposing an experimental design replicating specific activities. Should the results of this study be encouraging, a small scale replication with an experimental design will probably be proposed. Similar approaches (but thus far lacking the valuable component of randomized assignment of types or combinations of treatments) have been evolving in recent years in the Office of Education in such programs as Follow Through and the new Project Information Packages.[27]

Figure 4. A Proposed Experimental Design
Race Relations Emphasis

Remedial Program	Race Relations Emphasis: Yes	Race Relations Emphasis: No
present	I	II
absent	III	IV (Control Sample)

C. A PECULIARITY IN THE ANALYSIS OF COVARIANCE

To perform an analysis of variance means nothing more complicated than to compare achievement means of two groups of schools and perform a test of statistical significance to see if they differ from each other. In this study we found experimental schools to have mean achievement scores for black males about one-third grade higher than the control schools; in raw score units, about 15 points.

An analysis of covariance adds an additional wrinkle. It first statistically removes the effect of control variables (here called covariates) so that when we compare the experimental to the control group, we compare them only on that portion of the achievement test score which cannot be explained by background characteristics of the students. One reason why this is a good thing to do is that by removing the effects of background, we limit the range of possible scores that schools can obtain. While the actual raw school scores on the achievement test for black males range from 50 to 300, we know that the school which had a mean score of 50 is almost certain to have students from very poor families and conversely that the school with a score of 300 must serve a relatively wealthy attendance area. If we ask instead, by how many points do schools overachieve or underachieve compared to what we would expect given the social background of their students, we find the range not of 250 points but of only about 175 points. Against a range of 175 points, a 15 point difference is more impressive than it is in a world where scores range over 250 points. Thus the analysis of covariance, by limiting the range of scores, makes a difference between the experimental and control schools of a certain size more likely to be statistically significant. This indeed happened in our case.

The other reason why we might use an analysis of covariance is to adjust for differences in background characteristics of students in the two groups of schools. If the experimental school students are generally poorer than the control school students, we would expect their scores to be lower by a certain amount and this expected difference could be compared to the real difference. We did not intend to use

the analysis of covariance for this purpose because we expected the experimental schools and control schools being randomly selected to be identical or nearly so. In fact, the analysis of covariance indicated that the true difference between the experimental schools and the control schools was not 15 points but 25 points. The reason for this became obvious as soon as we looked at the tape; the experimental high schools generally had black students with lower socio-economic status than did the control schools.

How did this happen? It is possible that it is mere sampling error, but it is more likely that a bias was introduced in the process of recruiting schools to enter the experimental design. Many school districts refused to participate in the experimental design and their refusal, which came before any selection of experimental schools or control schools took place, should not have affected the design (except perhaps to limit the generalizability of the findings to the more cooperative school districts). However, a small number (18) of school districts did not refuse to participate until after they had been notified which schools in their districts had been designated as control schools. At that point York made what may have been a blunder. It would have been possible to permit these schools to withdraw from the experimental-control design but at the same time still study the schools in order to determine whether their withdrawal introduced a bias into the design. Instead, the schools were dropped. Did their withdrawal

introduce a bias? In the left hand portion of Figure 5 is a pair of rectangles representing the predominantly black experimental schools and control schools, each in their own rectangle; the two right hand rectangles represent predominantly white experimental and control schools. A horizontal line drawn across the two rectangles separates the districts at the top which agreed to cooperate in the study with the pairs of schools which were in districts which refused to cooperate after the experimental and control schools had been selected. Let us make the reasonable assumption that (subject to the vagaries of random sampling) the mean socioeconomic status of the full set of experimental schools was the same as the control schools. In order for the predominantly black control schools which were studied to have unusually high black social status and unusually low white social status, it must follow that the control schools which do not appear in the study because their superintendents refused to cooperate with the design must be the compensating schools with unusually low black social status and unusually high white social status. When we look at the right hand side of the figure, where we examine the predominantly white schools (which make up most of the sample), we find that the control schools which were studied have unusually high black social status although the white social status in these schools is no different from that of the experimental schools. It again follows that if the control schools that we studied have unusually high black status, the control schools which were lost to the study must have had unusually low social status. Thus, we

(Figure 5 about here)

Figure 5. Analysis of Bias in the Experimental Design

	Predominantly Black Schools		Predominantly White Schools	
	Experimental	Control	Experimental	Control
School Districts Studied: Known Difference	lo Bl. SES hi W. SES	hi Bl. SES lo W. SES	lo Bl. SES	hi Bl. SES
Withdrawals (assumed difference)	hi B. SES lo W. SES	lo B. SES hi W. SES	hi B. SES	lo B. SES
Mean SES, Total Sample	(assumed no difference)		(assumed no difference)	

present the hypothesis: superintendents were more likely to withdraw from the study after they learned which school was an experimental and which was control if they found that either high status white students in predominantly black schools had been deprived of the ESAP funds or if either white or black schools with unusually poor black students had been deprived of funds. If either of these conditions occurred the superintendent was more likely to find the experimental-control design unacceptable and withdraw. In retrospect this fits with our notion of the pressures that southern school superintendents are under, either from their own feelings about what they wanted to do or as a result of political pressure from their constituency. White children and especially high status white children in predominantly black schools are the hostages of a desegregation plan and the superintendent who wishes to remain in office had better pay very careful attention to them. News that he had permitted these children to be deprived of federal funds would have set very badly in the community. At the same time most superintendents are convinced that their most critical education problem is the low performance of poor blacks in desegrated while schools. Finding that they would be unable to use ESAP funds in order to provide remedial programs for these students would have inspired some superintendents to withdraw from the program. We admit that almost any pattern of differences between the experimental and control school could probably be explained by some plausible statement like this, but it does seem to us that the combination of poor black students who need help and rich white students whose parents have political power would represent

the most serious problems for a superintendent and would be the students whom he would be least willing to see denied ESAP funds.

As we said earlier, the simple solution would be to retain these schools in the study so that we could determine precisely what the social status of these students were. In this case, careful use of analysis of covariance made it possible to interpret the results and draw reasonably firm conclusions from the experimental design, but we should have anticipated the problem.

FOOTNOTES

[1] Southern Schools: An Evaluation of the Effects of the Emergency School Assistance Program and of School Desegregation. (Chicago: The National Opinion Research Center, October 1973). Final report to the Office of Planning, Budgeting and Evaluation of the Office of Education.

[2] Data compiled by the Office for Civil Rights, U.S. Department of Health, Education and Welfare. (1/1471 and 6/18/71 news releases).

[3] The figures are 11 per cent in the North and West and 9 per cent in the South. (1/13/72 news releases).

[4] RMC, Inc., Evaluation of the Emergency School Assistance Program, (Bethesda, Maryland: RMC, Inc., 1971).

[5] One school district requested funds to continue its chorus, on the grounds that it brought "harmony" between the races.

[6] American Friends Service Committee, Delta Ministry of the National Council of Churches, Lawyers Committee for Civil Rights Under Law, Lawyers Constitutional Defense Committee, NAACP Legal Defense and Education Fund, and the Washington Research Project, The Emergency School Assistance Program: An Evaluation, 1971.

[7] Donald T. Campbell and Julian C. Stanley, Experimental and Quasi-Experimental Design for Research, (Chicago, Illinois: Rand McNally, 1966).

[8] Donald T. Campbell and Albert Erlegacher, "How Regression Artifacts in Quasi-Experimental Evaluations can Mistakenly Make Compensatory Education Look Harmful," in J. Hellmuth, Ed., "Compensatory Education: A National Debate," Vol. 3, Disadvantaged Child, (New York, New York: Brunner/Mazel, 1970).

[9] John W. Evans and Jeffrey Schiller, "How Preoccupation with Possible Regression Artifacts Can Lead to a Faulty Strategy for the Evaluation of Social Action Programs: A Reply to Campbell and Erlebacher," In J. Hellmuth, Ed., "Compensatory Education: A National Debate," Vol. 3, Disadvantaged Child, (New York, New York: Brunner/Mazel, 1970).

[10] Eighteen of these did not withdraw until the actual randomization had been done and control students had been designated. These 18 districts cause a problem which we shall discuss later.

[11] Donald T. Campbell and Julian C. Stanley, Experimental and Quasi-Experimental Design for Research, (Chicago, Illinois: Rand McNally, 1966).

[12] One of the interesting problems of the sampling was the desire to sample at the time the school districts received their grants so the districts would know if they were in the sample and the identity of the control schools. Since grants were awarded in batches this made it necessary to project the number of grants to be awarded in the later batches in order to reach the desired sample size. This problem was, of course, compounded by the need to estimate refusal rates. Fortunately, both estimates turned out to be quite accurate.

[13] ESAA is Title VII of the Education Amendments of 1972 (Public Law 92-318).

[14] Preliminary reports in such areas as analysis plans and data collection management are different issues because they are part of planning for the total study.

[15] The same steps were followed in the regression analyses of program effects on attitudes toward integration except a rural-urban dichotomy was also introduced, producing eight grade-race-urbanism combinations.

[16] Student background included the percentages of students from broken homes, and measures of socioeconomic status (but not the usual ones; since 11 year olds don't know their parents' income, we used less conventional measures, such as whether their family had an air conditioner, or whether they had a bicycle).

[17] Had we not aggregated the data to the school level, Beta would have been an even less impressive +.04.

[18] See, for example, J.P. Guilford, Psychometric Methods, (New York, New York: John Wiley, 2nd edition, 1954) p. 400-401.

[19] We should point out that this effect, while both statistically and (we think) socially significant, still would look small using regression techniques; had we analyzed the experimental with regression the effect of ESAP would have been a standardized coefficient (B) of . 13.

[20] It is a significant note to the process of evaluation research that the origin of this analysis was with William Rock who was one of Goldberg's top staff members. Reviewing draft analysis plans for the study, Rock saw a gap and asked what we could learn about how to change teachers racial attitudes and behavior in desegregating schools.

[21] We have only one question in the questionnaire which is related to this issue, and that is a question addressed to teachers rather than principals. We asked the teachers "Do you feel you should let your students know how you feel about race relations or would this be improper?" Thirty-two per cent of the teachers felt that this would be improper.

[22] The Office of Education is organized such that several Deputy Commissioners report to the Commissioner. The Office of Planning, Budgeting and Evaluation reports to a Deputy Commissioner with no responsibilities for administering State or local educational grant or contract programs, thus helping to insure the independence of evaluation studies.

[23] In all, six different computers, in five different cities, were used on this project.

[24] News release for February 20, 1974.

[25] Using Coleman's terminology in James L. Coleman, "Policy Research in the Social Sciences", (Morristown, New Jersey" General Learning Press, 1972) p. 2.

[26] The report is available from the ERIC Document Reproduction Service in microficshe in over 500 college libraries and other education related institutions or be direct order in hard copy or microficshe. The ERIC identification numbers are ED 085426 for Volume 1 and ED 085427 for Volume 2. The survey instruments are printed as part of Volume 2. Orders may be placed either through ERIC, National Institute of Education, 1200 19th Street, N.W., Washington, D.C. 20036 or ERIC, P.O. Box 190, Arlington, Virginia 22210. Data tapes and documentation are available at cost from Mr. Patrick Bova, NCRC, 6030 South Ellis Ave., Chicago, Illinois 60637.

[27] G. Kasten Tallmadge, The Development of Project Information Packages for Effective Approaches in Compensatory Education, (Los Altos, California: RMC Research Corporation, 1974).